Human Immunology 76 (2015) 374–380

Contents lists available at ScienceDirect

journal homepage: www.elsevier.com/locate/humimm

Comparison of high-resolution human leukocyte antigen haplotype frequencies in different ethnic groups: Consequences of sampling fluctuation and haplotype frequency distribution tail truncation

Derek James Pappas [a], Alannah Tomich [b], Federico Garnier [c], Evelyne Marry [c], Pierre-Antoine Gourraud [c,d,*]

[a] Children's Hospital Research Institute, Oakland, CA 94609, USA
[b] University of California, Berkeley, Berkeley, CA, USA
[c] Registre France Greffe de Moelle, Agence de la Biomédecine, Paris, France
[d] University of California, San Francisco, San Francisco, CA, USA

ARTICLE INFO

Article history:
Received 6 June 2014
Accepted 21 January 2015
Available online 28 January 2015

Keywords:
Haplotype
Bone marrow donor registry
Population
Immunogenetic
Genetic epidemiology

ABSTRACT

High-resolution haplotype frequency estimations and descriptive metrics are becoming increasingly popular for accurately describing human leukocyte antigen diversity. In this study, we compared sample sets of publically available haplotype frequencies from different populations to characterize the consequences of unequal sample size on haplotype frequency estimation. We found that for low samples sizes (a few thousand), haplotype frequencies were overestimated, affecting all descriptive metrics of the underlying distribution, such as most frequent haplotype, the number of haplotypes, and the mean/median frequency. This overestimation was a result of random sample fluctuation and truncation of the tail end of the frequency distribution that comprises the least frequent haplotypes. Finally, we simulated balanced datasets through resampling and contrasted the disparities of descriptive metrics among equal and unequal datasets. This simulation resulted in the global description of the most frequent human leukocyte antigen haplotypes worldwide.

1. Introduction

The term "haplotype" was first coined by Ruggero Ceppellini in 1955 [1] to describe the co-occurrence of genetic polymorphisms of the HLA alleles along a single chromosome. Haplotypes have become the basic unit of analysis for studying humans as well as other organisms [2]. Haplotypes may be represented as blocks of DNA sequence variants that can be abstracted into allelic nomenclature at the level of a functional locus such as in the HLA system [3].

HLA genetic polymorphisms have been primarily studied in transplantation. As a pivotal factor in the transplantation outcome, HLA is one of the most preeminent genetic determinants in health care. Generally, HLA typing is performed for clinical purposes rather than for research purposes [4]. The immunogenetics community utilizes haplotype frequency data in proactive ways that go far beyond the scope of classical research interests in population and evolutionary genetics [5,6]. As such, haplotype frequencies are used for managing bone marrow donor registries [7–9] through prioritized lists of potentially matched donors that can be searched (Optimatch at ZKRD, Haplologic at NMDP, Easymatch in France), for targeting donor recruitment, optimal registry size computation, and optimal recruitment strategies, and for assessing cost of typing quality versus issue benefits [10–12].

Optimal study design and sample acquisition are dependent on the possibility of identifying haplotypes by segregation analysis of families or estimating haplotypes from population samples of phase-unknown unrelated individuals. The extreme polymorphisms of immunogenetic data present the following statistical modeling challenges: (1) the field is frequently plagued by very large sample sizes (>10,000 subjects); (2) low number of loci (approximately <20); (3) a large number of alleles per loci (approximately >50); and (4) high haplotype diversity (at least

Abbreviations: HLA, human leukocyte antigen; EM, expectation-maximization; LD, linkage disequilibrium; HSCT, hematopoietic stem cell transplantation.

* Corresponding author at: UCSF MS Genetics, 19A Building-NS235 UCBox-3206, Dept. of Neurology, School of Medicine University of California San Francisco, 675 Nelson Rising Lane, Mission Bay Campus San Francisco, CA 94158, USA. Fax: +1 415 476 5229.

E-mail address: pierreantoine.gourraud@ucsf.edu (P.-A. Gourraud).

http://dx.doi.org/10.1016/j.humimm.2015.01.029
0198-8859/

>1000). Advanced statistical methods are needed to provide reliable frequencies, although these are outside the scope of this work and have been studied previously regarding their general [13] and HLA-specific aspects [14,15]. Many haplotype frequencies have been published and represent a useful source of information describing the high diversity of HLA haplotype polymorphisms across populations [16,17]. Although much information is available on the dbMHC web site (http://www.ncbi.nlm.nih.gov/projects/gv/mhc/), this study reports the effort to provide a publically available collection of valuable HLA haplotype frequencies in an existing central repository. We investigated the role of sample size in the fluctuation of HLA haplotype frequency distributions, which may impact the results of studies using these estimations.

2. Materials and methods

2.1. Data sets

High-resolution HLA haplotype frequency datasets were identified and collected. Only those datasets in which the entire set of haplotype frequencies were made publically available [18–23] and that included a set of high-resolution HLA haplotype frequencies from the French Bone Marrow Donor Registry (France Greffe de Moelle) [24] were collected. We focused on high-resolution typing data of the HLA-A, HLA-B, and HLA-DRB1 alleles and on sample sizes exceeding 1000 (as of January 2012). Published data were made available through the Allele Frequency Net Database [25]. HLA haplotype frequencies were estimated using the expectation-maximization (EM) algorithms [26]. Investigation of the properties of the EM algorithm was not performed in this study because selected samples were used and because methods to account for all types of HLA typing results are still undergoing extensive discussion [15]. In this study, population refers to original datasets.

2.2. Population statistics

Several statistics were calculated to describe and compare haplotype frequencies between populations. These metrics included calculations of mean and median haplotype frequencies, the 25th and 75th percentiles of haplotype frequencies, the number of unique haplotypes, the sum of haplotype frequencies within a given range (e.g., the top 10), the minimum number of haplotypes that sum to a defined percentage (e.g., sum to 10%), and the number of haplotypes greater than or equal to a defined threshold (e.g., greater than 10%).

2.3. Random sampling of datasets

Effects of sample size were determined by taking 1000 random simulations from the French sample of haplotype distribution and generating samples with decreasing sizes for analysis. In each simulation, HLA-A, HLA-B, and HLA-DRB1 haplotype frequencies were assumed to be known because they have been generated from a set of published haplotypes with specified frequencies. These included sample sets of (N=) 100, 250, 500, 1000, 2000, 2500, 5000, 10,000, 15,000, and 20,000 individuals with 2N sets of haplotype. The statistics for each individual were recalculated and averaged across all individuals within each sample set. To make comparisons between different ethnic groups with equal sample sizes, the process was repeated with the datasets from each ethnic group and by setting the resampled population size to 1000 individuals. R software [27] was used to generate randomly sampled datasets and to calculate and graph all metrics.

3. Results

HLA frequency data were identified and collected from 13 datasets that spanned 11 populations, including European American, Asian American, and African American (Table 1). Table 1 demonstrates that for these datasets, several descriptive metrics differentiate the distributions of population haplotype frequencies. In all populations, there were only a few of the "most frequent" haplotypes (i.e., estimated frequency >1%). However, given the heterogeneity in sample sizes, direct comparisons cannot be made with confidence without understanding the effect of sample size on haplotype frequency estimations and distributions.

To test how sample size may affect the distribution of haplotype frequencies, the French haplotype distribution was sampled at random to create datasets of decreasing size and ranging from 100 to 20,000 individuals (Table 2). The same metrics as in Table 1 were compared across the reduced datasets, and all metrics demonstrated variable dependency based on sample size. Thus, variations in sample size affected the calculated metric. Log–log relations were observed for sample size relative to the number of haplotypes, the mean, the median, or any percentile cut-off computed from the distribution (Table 2 and Supplemental Fig. 1). Rectangular hyperbolic relations were observed between the log of the sample size and the log of either the sum of a given quantity of haplotypes or the minimum number of haplotypes that sum to a defined frequency threshold (Table 2 and Supplemental Fig. 2). The asymptotic nature of these log–log relationships suggests that for reasonably large samples (sizes >2000), differences in the sample size have increasingly smaller effects on the calculated metric. Finally, as the sample size decreased, the slope of the linear fit between the logged haplotype frequencies and their logged rank became increasing smaller (Fig. 1), demonstrating a log–log relation between the slope and sample size (Fig. 1, inset) (see Table 3).

Smaller sample sizes capture the most common haplotypes, and any increase in population size adds to the haplotype diversity; there were infrequent contributions from the rarest haplotypes. Consequently, the sample size affected the number of haplotypes whose frequencies were greater than a given threshold. On average, haplotype frequencies are overestimated in the smaller populations (Table 2 and Supplemental Fig. 3, red shaded region). In these datasets, smaller population sizes fail to capture the contribution from less common or rarer haplotypes; therefore, the frequency of these truncated alleles is redistributed to the more common haplotypes, resulting in an overestimation of their frequency. However, setting a threshold for comparison that is low compared with the sample size (e.g., the number of haplotypes whose frequency is greater than 0.01% in samples of less than 2000 individuals) can result in underestimation of haplotype frequency (Table 2 and Supplemental Fig. 3, blue shaded region).

To further refine how sampling fluctuation may affect the frequency estimation for individual haplotypes, the French dataset was resampled with N = 1000. The differences between the original haplotype frequency (reference) and the resampled haplotype frequency (sample) were plotted for each haplotype against the original haplotype frequency (Fig. 2). As expected, Fig. 2 shows that because of random sampling, frequencies may be overestimated (Fig. 2, red dots) or underestimated (Fig. 2, blue dots), and these deviations from the reference converged at discrete points given by the formula $x/2N$, where x corresponds to the haplotype count (e.g., 1, 2, 3) and N corresponds to the number of subjects within the sample set (2N is the number of haplotypes). The absolute minimum frequency for one haplotype that can be observed in 1000 individuals is 1/2000 or 0.05%. These minimum convergences exist at 0.05% (1/2000) and 0.1% (2/2000), whereas trends toward the minimum can be seen at 0.15% (3/2000) and 0.2% (4/2000).